# Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach

HHS Public Access

Author manuscript

*Nat Methods*. Author manuscript; available in PMC 2017 February 06.

Published in final edited form as:

*Nat Methods*. 2015 September ; 12(9): 823–826. doi:10.1038/nmeth.3473.

# Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach

**Raj Chari**[1,5], **Prashant Mali**[2,5,†], **Mark Moosburner**[3], and **George M. Church**[1,4,†]

[1]Department of Genetics, Harvard Medical School, Boston, MA, USA

[2]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA

[3]Scripps Institute of Oceanography, University of California San Diego, La Jolla, CA, USA

[4]Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, MA, USA

## Abstract

We develop an *in vivo* library-on-library methodology to simultaneously assess single guide RNA (sgRNA) activity across ~1,400 genomic loci. Assaying across multiple human cell types, end-processing enzymes, and two Cas9 orthologs, we unravel underlying nucleotide sequence and epigenetic parameters. Our results enable improved design of reagents, shed light on mechanisms of genome targeting, and provide a generalizable framework to study nucleic acid-nucleic acid interactions and biochemistry in high throughput.

RNA-guided genome engineering using the CRISPR/Cas system has enabled an unprecedented ability to perform site specific editing in a variety of genomes[1,2]. Several parameters affect this Cas9-guide RNA mediated genome targeting in mammalian cells. These include choice of Cas9 ortholog, spacer sequence composition, guide RNA (sgRNA) secondary structure, epigenetic status of target locus, Cas9-gRNA complex specificity, use of a double strand break versus nicking modality, and cell type intrinsic factors. A comprehensive analysis of these aspects will enable not just the ideal design of targeting reagents, but also shed light on the mechanisms that underlie genome targeting.

While studies have begun to reveal some of the rules in sequence composition relating to sgRNA activity[3,4], this has been limited to small numbers of loci in the genome and a single CRISPR/Cas9 system. In our study, we employ an *in vivo* and multiplex library-on-library



†Correspondence can be addressed to: Prashant Mali, pmali@ucsd.edu, George Church, gchurch@genetics.med.harvard.edu.
[5]These authors contributed equally

**ACCESSIONS**

All sequencing data generated in this study are deposited in the NCBI SRA under accession number SRP048540.

**AUTHOR CONTRIBUTIONS**

R.C., P.M. designed the study and performed the experiments. R.C., P.M. wrote and edited the manuscript. All authors approved the final version of the manuscript. R.C. implemented custom python software and performed data analysis. M.M. provided technical assistance. G.M.C. supervised the project.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

approach to assess sgRNA activity across ~1,400 genes for CRISPR/Cas systems from multiple bacterial species, both in the presence and absence of multiple end-processing enzymes. Specifically, we synthesize a library of sgRNAs and evaluate their activity against a library of targets. To enable simultaneous assessment across all loci, we perform two independent experiments – in the first, we synthesize the corresponding guide RNA targets into lentiviruses and transduce a population of cells at a high titer (Fig. 1A). These synthesized targets bear shared flanking sequences, enabling analysis of all synthesized target loci via a single PCR followed by high throughput sequencing. In the second method, we do a targeted pull down of corresponding endogenous loci and do high throughput sequencing to assay non-homologous end joining (NHEJ) profiles (Fig. 1B). Both assays enable simultaneous readout across all target loci, however the results of the former assay corresponds to raw targeting rates (since the targets are integrated in expressed lentivirus inserts), the latter results are modulated by underlying epigenetic status of corresponding loci. Together, the assay thus enables a comprehensive assessment of Cas9-sgRNA mediated genome editing.

First, we performed both biological and technical replicates of our experiments to assess how reproducible our library vs. library approach is for the CRISPR/Cas9 system from *Streptococcus pyogenes* ($Cas9_{Sp}$). Selecting one representative pair of biological replicates, we observed a correlation of 0.84 between the rates observed in the first and second biological replicate (Fig. 2A, Supplementary Fig. 1). As a separate control, we performed the experiment with a nickase version of $Cas9_{Sp}$ and observed an expected drastic reduction in NHEJ with $Cas9_{Sp}$ nickase vs. $Cas9_{Sp}$ nuclease (Fig. 2B). We then employed this method in an additional cell line (K562) (Fig. 2C) as well with another Cas9 ortholog from *Streptococcus thermophilus* ($Cas9_{St1}$). We observed that the mutation rates for $Cas9_{St1}$ were lower compared to $Cas9_{Sp}$ (Fig. 2D).

The co-expression of end-processing enzymes along with endonucleases had recently been shown to increase NHEJ-associated mutation frequency[5]. However, these experiments were performed over a small number of loci. By utilizing this library-on-library approach in the context of these end-processing enzymes, we can perform an unprecedented level of analyses to assess the impact of these enzymes over a large and diverse set of sequences. Of the four different end-processing enzymes we tested, we observed increased mutagenesis rates associated with TREX2 (Fig. 2E, Supplementary Fig. 2), largely attributed to deletions (Supplementary Fig. 3) and to lesser degree, Artemis (Supplementary Fig. 4), as well as an impact on the size of NHEJ-induced deletions observed (Supplementary Fig. 5). In a separate experiment, however, we observed that TREX2 also increases off-target mutation rates (Supplementary Fig. 6), suggesting careful sgRNA design is imperative if using TREX2 as a means to increase on-target mutagenesis rates.

We next sought to determine whether specific rules or motifs were enriched or lost in the sgRNAs which demonstrated high activity compared to those with little or no activity. For both Cas9s, we identified the sgRNA sequences in the top quartile in each experiment, both in the presence and absence of our four end processing enzymes. Taking the overlap of the top quartile and bottom quartile in the five experiments, we identified 133 high activity sgRNAs and 146 in low activity for $Cas9_{Sp}$ and 82 and 69 for $Cas9_{St1}$ (Supplementary Data

1). For each set, we compiled nucleotide frequencies at each position and then compared those frequencies between the high and low activity sets. The most drastic difference observed was at position 20, next to the protospacer adjacent motif (PAM), where there was a strong preference for 'G' and low preference for 'T' (Fig. 3A). While some of these features have been observed previously, we additionally note that the 'G' in position 20 is important to both $Cas9_{Sp}$ and $Cas9_{St1}$ (Supplementary Fig. 7), suggesting that this feature may be general to CRISPR/Cas9 systems.

In order to capture higher order relationships between the high and low activity sgRNAs, we generated a support vector machine model. Ten-fold cross validation of our models achieved an average accuracy of 73.2% for $Cas9_{Sp}$ and 81.5% for $Cas9_{St1}$. We then assessed the accuracy of the predictions of the SVM experimentally. For each Cas9, we selected ten previously untested sgRNAs of which five were predicted to have high activity and five predicted to have low activity and assessed them on an individual basis across a set of seven diverse cell lines (Supplementary Table 1). While the difference in mutagenesis rates between predicted high and low sgRNAs is noticeable for $Cas9_{Sp}$, the difference is even more striking for $Cas9_{St1}$. Furthermore, this trend was observed across multiple cell types, suggesting our model is generalizable (Supplementary Fig. 8 and 9).

A recent study using an sgRNA library of a similar size to this study, targeting numerous sites within nine endogenous genes, had also deciphered rules which may govern sgRNA activity[3]. To assess the quality of our SVM classifier, we scored the 1,841 sgRNA sequences used in their study and correlated our scores with their predicted scores and data. We observe a modest correlation between our predicted scores with their data (Supplementary Fig. 10). It is somewhat expected that there would be some variability as: (1) the set of sgRNA sequences were different to build the model, (2) we assess impact within 72 h as opposed to two weeks which may limit selection bias, and (3) our study uses a completely sequence-based (and hence sgRNA activity) readout, while the other study employed a phenotype-based readout.

Due to the ability of lentiviral integration to generally promote states of open chromatin, the employment of an integrated target library allows for an interrogation of sgRNA activity as purely based on sequence composition as possible. However, when targeting endogenous loci themselves, this is typically not the case. In a parallel experiment, we transfected our $Cas9_{Sp}$ and $Cas9_{St1}$ sgRNA libraries in target-naïve 293T cells, enriched for sequence surrounding all of our target sites, and performed high throughput sequencing. Unsurprisingly, the overall mutagenesis rates were markedly less than what we had observed using our integrated target site library for both Cas9s (Supplementary Fig. 11).

Previous studies utilizing $Cas9_{Sp}$ immunoprecipitation experiments have observed binding preferences of $Cas9_{Sp}$ to areas of higher DNA accessibility in the context of single sgRNAs[6,7]. To evaluate this metric in our system, we obtained DNase I hypersensitivity (DHS) data for the HEK293T cell line from ENCODE[8] and compiled DNase-seq values for each site. Taking the top quartile of sites with the highest % of NHEJ observed and bottom quartile of sites with the lowest %, we compared the DNase-seq values between the groups and observed a higher range of values in the regions where we saw higher activity as

compared to those with low activity (Fig. 3B). In addition, examining H3K4-trimethylation status, a histone mark associated with actively transcribed genes[9], we also see a similar statistically significant enrichment (Fig. 3C, Supplementary Fig. 6). Intriguingly, we observed a small set of outliers in the group of sites with low mutagenesis rates, which showed high DNase-seq values, and when we subsequently employed our classifier on the corresponding sgRNA sequences, our classifier determined that 76.7% (23/30) of these sequences would be considered poor (Fig. 3D). Taking together, it suggests both locus accessibility and sequence composition of the sgRNA are important in determining sgRNA activity.

We then sought to directly compare the mutagenesis rates at the lentiviral target sites with the endogenous sites. Using those sites which had sufficient coverage in both sets of experiments, we found a strong correlation ($r = 0.422$, $P = 1.6 \times 10^{-53}$). Given the endogenous dataset is comprised of sites with heterogeneous epigenetic profiles while the lentiviral dataset represents a relatively more homogeneous set, this correlation is not surprising. Taking the DNase I values calculated above, we sorted the endogenous mutation rate dataset from most to least accessible and performed correlations of the top 100, 200, up to 1000 most accessible sites to account for the accessibility. As we enrich for more accessible sites, our correlations continually improve (Supplementary Fig. 12), suggesting that lentiviral target sites are highly accessible and our parallel lentiviral scheme is important to extract the true underlying sequence features associated with sgRNA activity.

Off-target activity of the CRISPR/Cas9 system has generally been recognized as an issue that has needed to be addressed prior to any extensive use in gene therapy. A number of studies have assessed issues related to specificity; from identifying the breadth of the problem to solutions[10–17]. Using a recently published off-target dataset study[17], we examined the relationship between specificity and activity and observed no tangible relation (Fig. 3E). This result highlights a need to account for both specificity and activity in the sgRNA design process. To this end, we have compiled lists of human and mouse exome-wide $Cas9_{Sp}$ and $Cas9_{St1}$ target sites which are predicted to be both highly active and specific, determined by the *CasFinder* algorithm (Supplementary Data 2 to 5)[18]. We also have made available a software package which can identify and/or score sgRNA sites using user-defined sequences as input (Supplementary Software 1).

In summary, we have employed an *in vivo* and multiplex library-on-library approach to assess sgRNA activity across thousands of loci. Assaying also across multiple human cell types, Cas9 orthologs, and end-processing enzymes we unravel underlying nucleotide sequence and epigenetic parameters and define a predictive model for sgRNA activity. We demonstrate that our *in vivo* library-on-library approach is highly tractable and can be extended to newly identified CRISPR/Cas systems as well as capturing other nucleic acid/nucleic acid interactions in high throughput.

# ONLINE METHODS

## Selection of genes and target sites

A list of genes that would be considered of high value, encompassing ion channels, receptors and genes in the cancer gene census[1] was first derived. Next, using the hg19 RefSeq annotations for each gene, the exon sequences with 75 nucleotides of flanking sequence were downloaded from the UCSC Table Browser[2]. Custom python scripts were written to identify all unique Cas9 *S. pyogenes* ($Cas9_{Sp}$, $N_{20}NGG$) and *S. thermophilus* ($Cas9_{St1}$, $N_{20}NNAGAAW$) sites with the exons. These sites were then aligned against entire Hg19 genome sequence using *SeqMap*[3] and the sites 1) with no three-nucleotide off-targets in the genome and 2) targeting the 5' most exon were retained. In total, sites were successfully generated for 1,362 genes for $Cas9_{Sp}$ (Supplementary Data 6) and 1,449 for $Cas9_{St1}$ (Supplementary Data 7), with one site per gene. Sequencing of the original plasmid libraries revealed fairly uniform representation of both targets and sgRNAs (Supplementary Fig. 13). For the target sites, 1,344/1,362 of the $Cas9_{Sp}$ target sites and 1,276/1,449 of the $Cas9_{St1}$ were detected from this analysis. For the sgRNA libraries, we were able to detect 100% of the sgRNA sequences that we had synthesized for both Cas9s from the sequencing analysis.

## Target library synthesis

For each target site, we synthesized the target site as part of larger sequence in a 170 base long oligonucleotide using an in house CustomArray machine. Within each 170 base sequence, the flanking 25 base sequence was a sequence orthogonal to the human genome as defined previously used for efficient PCR amplification[4]. The adjacent ten nucleotides to each primer were given a unique 10 base barcode sequence which was at least 2 units in hamming distance away from any other 10 base barcode used. Finally, for the remaining 100 bases, the center of this sequence encompassed the 23 base target site (27 bases for $Cas9_{St1}$) and 39 and 38 bases of endogenous flanking sequence on each side for the $Cas9_{Sp}$ target site and 37 and 36 bases of sequence for $Cas9_{St1}$ (Supplementary Fig. 14). Sequences were amplified and cloned into a lentivector (Addgene #26777) for subsequent integration experiments. Plasmid libraries were sequenced and assessed for mutations introduced due to synthesis errors. These mutations were subsequently used to filter out synthesis related errors from bona fide mutations observed in experimental samples. Synthesized oligonucleotide sequences are shown in Supplementary Data 8 ($Cas9_{Sp}$) and 9 ($Cas9_{St1}$).

## sgRNA library synthesis

For each target site generated, the corresponding sgRNAs were synthesized on CustomArray at the same length as the target site (Supplementary Fig. 14). Additional flanking scaffold sequence was added to the synthesized protospacer to enable Gibson assembly mediated cloning into the destination vector (Addgene #24150). The general methodology and scaffolds used for $Cas9_{Sp}$ and $Cas9_{St1}$ were as previously described[5,6]. sgRNA libraries have been deposited in Addgene.

### Library-on-library experiments

Cells were first transduced with a high titer of the lentivirus target library to create a pool of cells bearing the target library. Next, enrichment of successful target integration was performed using puromycin selection (2 μg/mL). These transduced cells ($1\times10^6$) were then transfected with Cas9 (5 μg), corresponding gRNA library (5 μg), and end processing enzyme/empty vector (5 μg). Plasmids encoding TREX2 (#40210), Artemis (#40211), and empty backbone (#39991) were obtained from Addgene. Given the high probability of a single cell achieving multiple NHEJ events due to the presence of multiple target sites, DNA from cells are harvested 72hrs post-transfection to minimize loss of cells with large amounts of nuclease activity. For 293Ts, lipofection (using Lipofectamine 2000, Invitrogen) was used; and for K562s, nucleofection using the 4D nucleofector (Lonza) was used to deliver the DNA. Cell lines used were obtained from ATCC and tested for mycoplasma.

### High throughput sequencing library preparation

DNA was extracted using Qiagen DNeasy kit with RNAse treatment. In order to add the appropriate adapters for high throughput sequencing, libraries were prepared in two consecutive PCR reactions. The first PCR to retrieve the site from genomic DNA and the second PCR to add the appropriate adapter sequences for Illumina sequencing.

To ensure minimal variability due to DNA sampling, 8 to 10 100 μL PCR reactions were performed for each sample using 1 μg of DNA in each PCR reaction. Briefly, in each reaction, 50 μL of KAPA 2× HiFi ready mix with 3 μL of each primer and 1.2 μL of SYBR green were added together with the appropriate amount of water and template to total 100 μL. Quantitative real-time PCR was used to ensure libraries were not amplified past the linear amplification phase with PCR reactions terminated accordingly. For each sample, multiple PCR reactions were pooled and then PCR purified with Qiagen columns and then eluted in 40 μL of elution buffer. Due to the presence of primer dimers, half of each eluate was run on an Invitrogen 2% E-gel EX for 10mins and then subsequently gel purified and extracted using Qiagen columns. In parallel, primers designed to capture the sgRNA sequences were also employed and in this case, since just the prevalence of each guide was needed, only 1 – 100 μL reaction was used for each sample.

For the second PCR step, 6 – 25 μL PCR reactions were set up for sample. In this case, 12.5 μL of KAPA 2× HiFi ready mix, 0.75 μL of each primer and 0.3 μL of SYBR green were added together. Real-time qPCR was once again utilized to ensure libraries were not amplified for too long and in this case, no more than 7 cycles were performed. PCR reactions were pooled and purified using Qiagen columns and these purified PCRs were also gel purified as well. Quantification of DNA was performed using Qubit with the high sensitivity assay. Equal amounts of each sample were then pooled into one tube and sequenced on an Illumina HiSeq 2500 on the Rapid Run Mode with paired-end 150 bp reads at the Biopolymers Facility in the department of Genetics at Harvard Medical School. For the K562 control and treated sample, libraries were prepared the exact same way but were sequenced on an Illumina MiSeq at the Molecular Biology Core Facility at Dana Farber Cancer Institute using the paired end 150 bp mode.

For the custom capture sequencing of endogenous targets, probes were designed flanking 200 bp on each of the target site for Agilent Sure Select and synthesized by the Beijing Genome Institute (BGI). Extracted DNA was then sent to Hong Kong for capture enrichment and sequencing on one lane in an Illumina HiSeq. All data have been deposited in the sequence read archive (SRA) under accession number SRP048540.

### Illumina sequencing data processing – target site analysis

FASTQ files were first de-segregated into each sample using in-house developed python scripts. For each sample, each set of read pairs were first merged using FLASH[7] into one larger contig and subsequently aligned to the custom reference sequences that were originally designed using BWA[8]. Next, alignments were filtered for uniqueness (no other alignments to any other sequence) and for length (136 bp). 136 bp was chosen as this is the minimum size guaranteed to cover the entire 100 bp payload sequence of our target site. Finally, samtools was then used to convert SAM to BAM files and generate pileups[9]. The computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.

### Illumina sequencing data processing – endogenous site analysis

FASTQ files obtained from BGI were then mapped to the entire Hg19 genome using bwa mem in the BWA package[8]. Reads were filtered to those that were successfully mate paired as well as not having multiple hits in the genome. Similar to above, SAM tools was used to obtain pileups from the generated SAM files[9].

### Calling NHEJ associated mutations

Given the prevalence of errors in synthesized oligonucleotides from the custom array, the original target-site plasmid library was sequenced and assessed for insertions and deletions (indels). These indels and those found in the untreated sample were subsequently used to filter indels observed in the treated samples to ensure indels called were truly due to the Cas9 and sgRNAs. Custom python scripts (available upon request) were used to analyze pileups to generate list of mutated sequences per site, the total # of mutations observed, the types of mutations observed and the total coverage at each. Single base substitutions were excluded from the analysis due to the high degree of false positives. Only mutations which spanned any part of the target site were considered. A minimum of 100 mapped reads at a given site was required in order to be considered for mutational analysis. Box plots and scatter plots were produced using the matplotlib[10]. Raw mutation rates observed in 293T cells for $Cas9_{Sp}$ nuclease, $Cas9_{Sp}$ nickase, and $Cas9_{St1}$ are provided in supplementary datasets (Supplementary Data 10–12). Rates observed in K562 with $Cas9_{Sp}$ nuclease are provided in Supplementary Data 13.

### Sequence motif analysis and support vector machine (SVM) model generation

For each Cas9, the base distributions between the high activity and low activity sgRNA sequences were compared using a 2 × 4 Fisher's exact test in R. Five base pairs of sequence on each side of the target site were also analyzed. Since bases at position 22 and 23 for $Cas9_{Sp}$ and positions 23, 24, 25 and 26 for $Cas9_{St1}$ are fixed, these positions were excluded

in the calculation. P-values generated were corrected for multiple hypothesis testing using the Benajmini-Hochberg formula.

To build the SVM classifier model, the same sets of sequences were encoded using a 4-bit binary scheme. The model was generated using SVM$^{Light\ 11}$. To generate the support vector machine (SVM) classifier, the high activity group of sgRNAs were categorized as “+1” and the low activity group “−1”. For each 23 bp sequence (27 bp for Cas9$_{St1}$), we encoded each character using a 4-bit binary system. For “A”, the encoding was '0001'; for “C” it was '0010'; for T it was '0100' and for “G”, it was '1000'. This was done to ensure the distance between any two bases would be equal. Each string of ‘1’s and ‘0’s, one per sequence, were then used as the input for the SVM. Ten-fold cross validation was used to assess the classifier and then the entire set was used to generate the model that was employed on the list of highly specific sgRNAs obtained from CASFinder[12].

## Comparison of the derived SVM model with previously published literature

A supplementary table detailing the observed sgRNA activity and predicted score for 1,841 sgRNA sequences was obtained from a recently published study[13]. Sequences were classified using our model and raw SVM prediction scores were determined. These scores were plotted against the predicted score by the published model as well as observed sgRNA activity and a Pearson correlation coefficient was calculated using scipy. Plots were done using matplotlib.

## Epigenetic data analysis

DNase-seq (GSM1008573) and H3K4-trimethylation (GSM945288) data, generated as part of the ENCODE consortium[14], were downloaded from the UCSC Genome Browser[15]. For each site, 225 bp of flanking sequence were used on each side, totaling a region of 473 bp in size for Cas9$_{Sp}$ and 477 bp in size for Cas9$_{St1}$. Data for each region were extracted using the bigWigAverageOverBed tool[16]. Distributions of values were compared using a t-test with an unequal variance assumption. The stats module in scipy was used to perform the t-test.

## Analyses of specificity and activity

To examine the relationship between specificity and activity, a recently published study whereby a genome-wide analysis of double stranded DNA breaks in the presence of Cas9 and an sgRNA was used[17]. For each observed site, the target sequence was scored with our SVM classifier and a scatter plot was generated comparing the SVM scores vs. the number of reads obtained in the GUIDE-Seq study.

For the lists of target sites in Supplementary Data 2 and 3 corresponding to Cas9$_{Sp}$ sites in human and mouse, they were obtained from http://arep.med.harvard.edu/CasFinder/. Next, the list of 2,712,189 sites identified previously for human and 2,733,854 sites for mouse were processed through the Cas9$_{Sp}$ SVM classifier, prediction scores were determined using SVM$^{Light\ 11}$, and a distribution of these scores was created. Finally, each site generated by CASFinder[4] was then scored with the SVM and its percentile rank in distribution of scores was calculated and reported.

For Supplementary Data 4 and 5, since the PAM for $Cas9_{St1}$ was different from the one used by *CASFinder* initially, *CASValue* (part of *CASFinder*) was re-run with default parameters using a PAM of "NNAGAAW" on the list of 376,758 sites for human and 350,497 for mouse. In total, *CASValue* deemed 195,861 human sites and 209,858 mouse sites as highly specific. Similar to $Cas9_{Sp}$, a score distribution was generated from the larger list sites and the percentile rank of the CASValue site was calculated and reported.

## Analysis of TREX2 on off-target mutagenesis

Three sgRNAs with known off-target sites were used for this analysis[14]. For each on-target site, three off-target sites were assessed for NHEJ-induced mutagenesis. sgRNAs were cloned into the pLKO.1 backbone. Primer sequences and target sites are listed in Supplementary Table 2. 500,000 293T cells were seeded in each well of a 6-well plate. Approximately 24hrs later, 2 μg of sgRNA and 2 μg of $Cas9_{Sp}$ were co-transfected using Lipofectamine 3000 at a ratio of 2:1. For the wells which included TREX2, 2 μg of pExodus-TREX2 was transfected as well. Seventy-two hours post-transfection, cells were harvested and DNA was extracted with 200 μL of Quick Extract solution.

Five μL of extracted DNA were then used as template in an 100 μL KAPA HiFi reaction and target sequences were amplified using quantitative PCR. PCR products were then purified using SPRI-bead purification and subsequently used as template for a $2^{nd}$ round of PCR to add Illumina adapter sequences. Final products were ran on a 2% E-gel EX and then extracted and purified using the Zymoclean Gel DNA recovery kit. All final PCR products were barcoded and pooled for Illumina sequencing analysis on the Illumina MiSeq. Paired-end sequencing data generated from the MiSeq were first merged using FLASH and then mapped to amplicon sequences using BWA-MEM. Pileups were generated using samtools and NHEJ-induced mutagenesis was determined using custom python scripts. To minimize the influence of Illumina error-rates, single base substitutions were excluded as only insertion and deletion events were included.

## Validation of predicted sgRNA activity

In total, 20 independent target sites were selected for individual sgRNA targeting. The 20 sites corresponded to 5 sites predicted for high activity and 5 sites predicted for low activity for both $Cas9_{Sp}$ and $Cas9_{St1}$ which were obtained from Supplementary Data 2 and 4, respectively. sgRNAs were cloned into the pLKO.1 backbone. Target site sequences and primers used amplify regions encompassing target sites are listed in Supplementary Data 11.

Similar to the TREX2 experiments, 500,000 293T cells were seeded and approximately 24hrs later, 2 μg of sgRNA and 2 μg of $Cas9_{Sp}$ (or 2 μg of $Cas9_{St1}$) were co-transfected using Lipofectamine 3000 at a ratio of 2:1 and DNA was harvested 72hrs post-transfection. DNA extraction, library preparation, and data analysis were performed exactly the same as it was done for the TREX2 experiments.

For experiments in A549, U2OS, HepG2 and SK-NAS cell lines, cells were seeded in a 24-well plate and for each sgRNA, 500 ng of sgRNA and 500 ng of Cas9 plasmid were co-transfected using Lipofectamine 3000 at a ratio of 2:1 and DNA was harvested 72hrs post-

transfection. For K562 and PGP1 induced pluripotent stem (iPS) cells, DNA was transfected using the Lonza 4D nucleofector.

### Code availability

All custom python scripts used in this study are available upon request.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We would like to acknowledge J. Aach for help with CASFinder and useful discussion, B. Turczyk with custom array oligonucleotide synthesis, S. Byrne for providing PGP1 induced pluripotent stem cells, and A. Chavez for useful discussion. This work was supported by NIH grant P50 HG005550. R.C. was supported by a Banting Fellowship from the Canadian Institutes of Health Research. P.M. is supported by UCSD startup funds and a Burroughs Wellcome Career Award at the Scientific Interface.

## MAIN REFERENCES

1. Mali P, et al. Science. 2013; 339:823–826. [PubMed: 23287722]
2. Cong L, et al. Science. 2013; 339:819–823. [PubMed: 23287718]
3. Doench JG, et al. Nat. Biotechnol. 2014; 32:1262–1267. [PubMed: 25184501]
4. Gagnon JA, et al. PLoS ONE. 2014; 9:e98186. [PubMed: 24873830]
5. Certo MT, et al. Nat. Methods. 2012; 9:973–975. [PubMed: 22941364]
6. Kuscu C, Arslan S, Singh R, Thorpe J, Adli M. Nat. Biotechnol. 2014; 32:677–683. [PubMed: 24837660]
7. Wu X, et al. Nat. Biotechnol. 2014; 32:670–676. [PubMed: 24752079]
8. ENCODE Project Consortium. Nature. 2012; 489:57–74. [PubMed: 22955616]
9. Koch CM, et al. Genome Res. 2007; 17:691–707. [PubMed: 17567990]
10. Ran FA, et al. Cell. 2013; 154:1380–1389. [PubMed: 23992846]
11. Mali P, et al. Nat. Biotechnol. 2013; 31:833–838. [PubMed: 23907171]
12. Tsai SQ, et al. Nat. Biotechnol. 2014; 32:569–576. [PubMed: 24770325]
13. Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK. Nat. Biotechnol. 2014; 32:279–284. [PubMed: 24463574]
14. Fu Y, et al. Nat. Biotechnol. 2013; 31:822–826. [PubMed: 23792628]
15. Guilinger JP, Thompson DB, Liu DR. Nat. Biotechnol. 2014; 32:577–582. [PubMed: 24770324]
16. Pattanayak V, et al. Nat. Biotechnol. 2013; 31:839–843. [PubMed: 23934178]
17. Tsai SQ, et al. Nat. Biotechnol. 2015; 33:187–197. [PubMed: 25513782]
18. Aach J, Mali P, Church GM. bioRxiv. 2014

## ONLINE REFERENCES

1. Futreal PA, et al. Nat. Rev. Cancer. 2004; 4:177–183. [PubMed: 14993899]
2. Karolchik D, et al. Nucleic Acids Res. 2004; 32:D493–D496. [PubMed: 14681465]
3. Jiang H, Wong WH. Bioinforma. Oxf. Engl. 2008; 24:2395–2396.
4. Xu Q, Schlabach MR, Hannon GJ, Elledge SJ. Proc. Natl. Acad. Sci. U. S. A. 2009; 106:2289–2294. [PubMed: 19171886]
5. Esvelt KM, et al. Nat. Methods. 2013; 10:1116–1121. [PubMed: 24076762]
6. Mali P, et al. Science. 2013; 339:823–826. [PubMed: 23287722]
7. Magoč T, Salzberg SL. Bioinforma. Oxf. Engl. 2011; 27:2957–2963.

8. Li H, Durbin R. Bioinforma. Oxf. Engl. 2009; 25:1754–1760.
9. Li H, et al. Bioinforma. Oxf. Engl. 2009; 25:2078–2079.
10. Michael, Droettboom, et al. 2014
11. Schölkopf, B., Burges, CJC., Smola, AJ. Advances in kernel methods: support vector learning. MIT Press; 1999.
12. Aach J, Mali P, Church GM. bioRxiv. 2014
13. Doench JG, et al. Nat. Biotechnol. 2014; 32:1262–1267. [PubMed: 25184501]
14. ENCODE Project Consortium. Nature. 2012; 489:57–74. [PubMed: 22955616]
15. Karolchik D, et al. Nucleic Acids Res. 2014; 42:D764–D770. [PubMed: 24270787]
16. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. Bioinforma. Oxf. Engl. 2010; 26:2204–2207.
17. Tsai SQ, et al. Nat. Biotechnol. 2015; 33:187–197. [PubMed: 25513782]

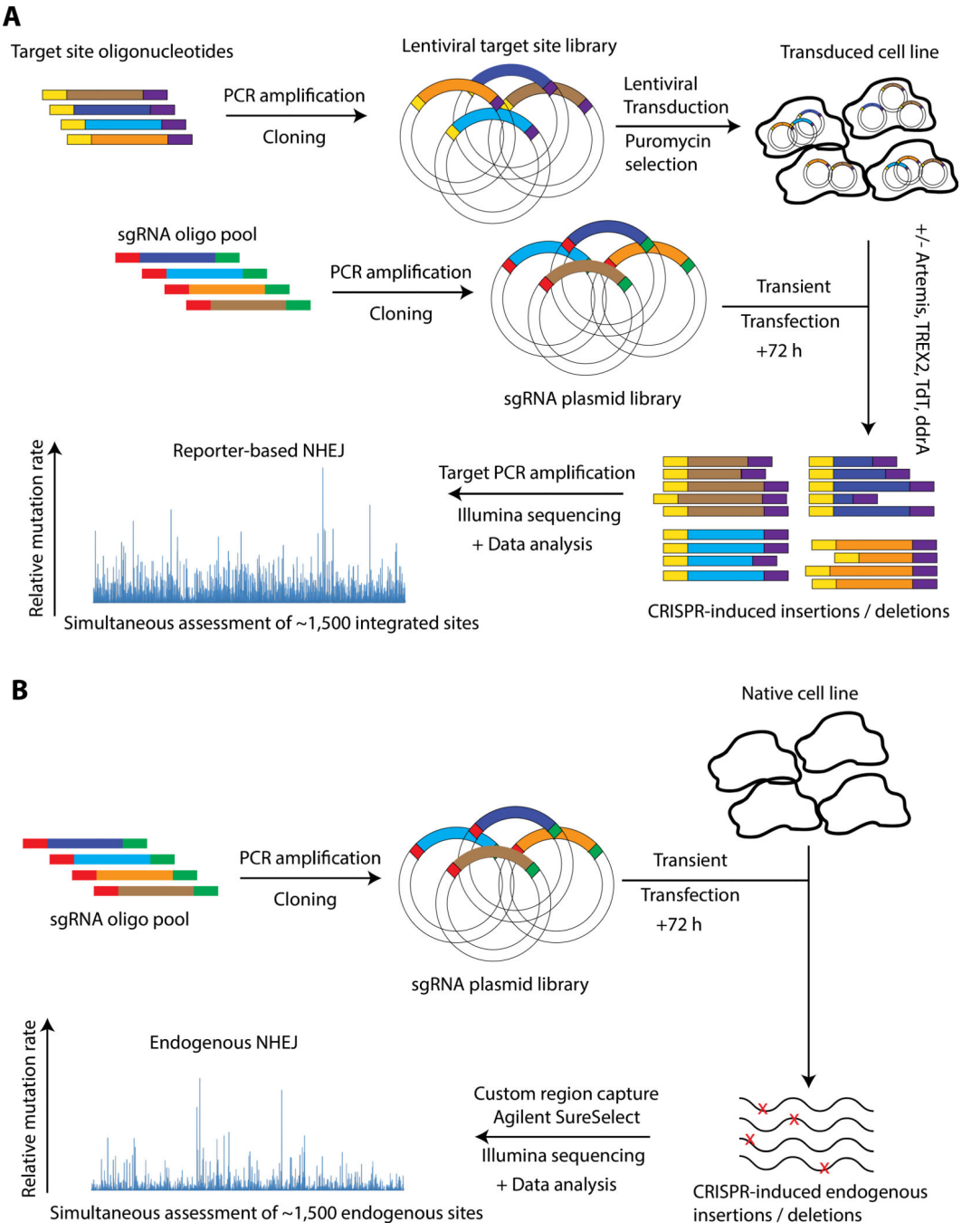


**Figure 1.**

Schematic of the library-on-library approach employed in our study. **(A)** sgRNA sequences corresponding to ~1,400 endogenous target sites were synthesized and cloned to make a sgRNA plasmid library. In parallel, replicas of the ~1,400 target sites were synthesized and cloned to make a lentiviral plasmid library. Subsequently, this lentiviral target library was integrated in our cells. Next, the sgRNA plasmid library was transiently transfected and cells were harvested for DNA 72 h post-transfection. Primer sequences designed against the constant sequences surrounding the target sites were used for PCR and libraries were

prepared for Illumina sequencing. **(B)** In target site naïve cells, the above mentioned sgRNA plasmid library was transiently transfected and DNA was harvested 72 h post-transfection. Agilent SureSelect enrichment was performed on the genomic DNA with a custom set of probes specific to regions encompassing the ~1,400 target sites and libraries were prepared for Illumina sequencing.

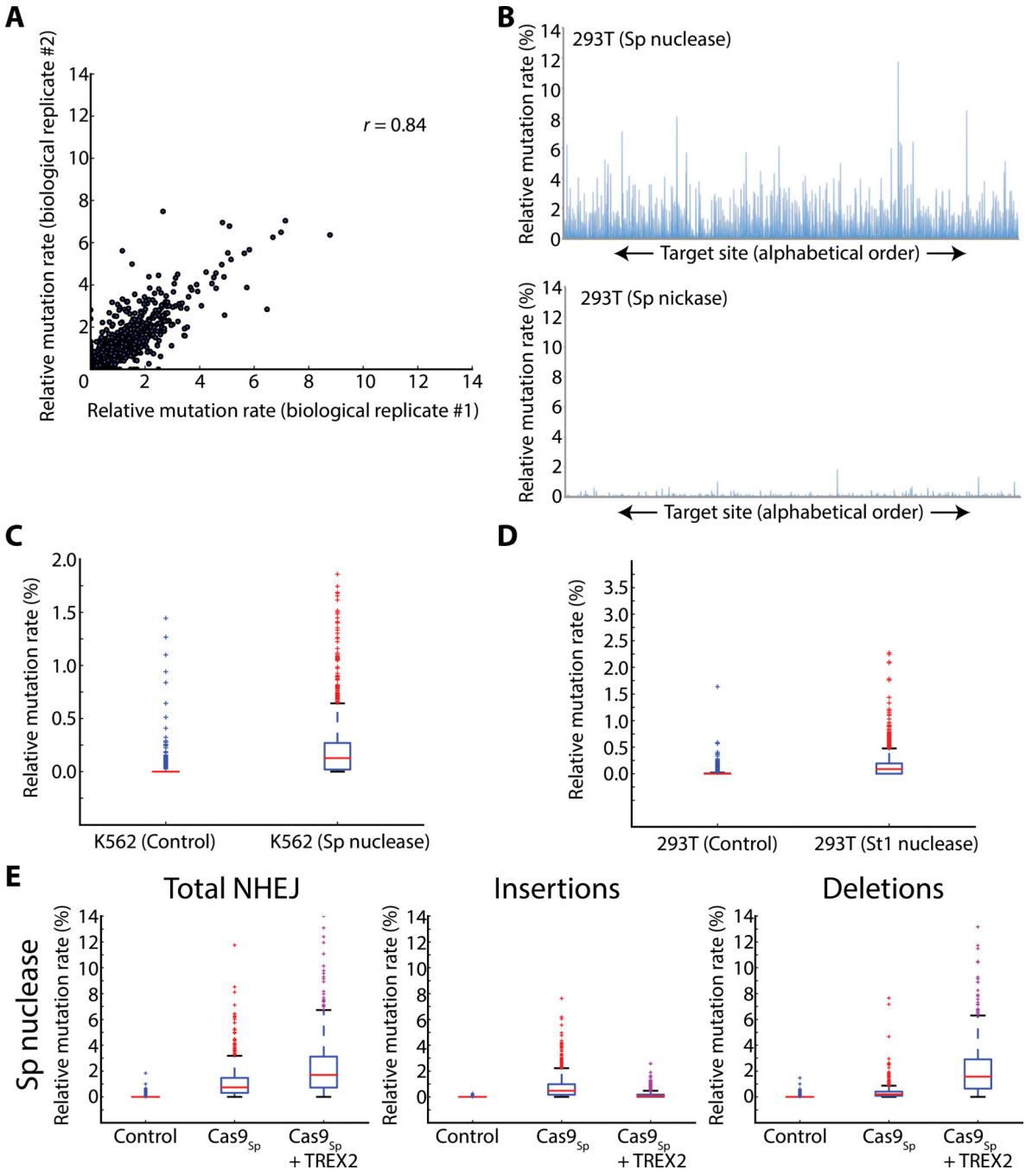


**Figure 2.**

Versatility of the library-on-library approach. **(A)** Scatter plot assessing the reproducibility of observed relative mutation rates across biological replicates. Data compared were from two independent sgRNA library transfections. **(B)** Comparison of activity between $Cas9_{Sp}$ nuclease vs. $Cas9_{Sp}$ nickase across all of the assessed sites. Box plots depicting the range of activities for **(C)** $Cas9_{Sp}$ nuclease in K562 cells. In total, 1,206 sites are represented for the control and 1,228 for the Cas9Sp nuclease treated cells. **(D)** $Cas9_{St1}$ nuclease in 293T cells. In total, data for 1,172 sites are shown for the control and 1,169 for the nuclease treated

cells. **(E)** Impact of TREX2 on altering patterns of NHEJ. A heavy bias towards deletions and slightly away from insertions is observed upon addition of TREX2. Data shown is for 293T using $Cas9_{Sp}$ nuclease. For all box plots, the size of the boxes represent the interquartile range (IQR) of the data. Whiskers are drawn to 1.5 * IQR. Red horizontal line represents the median. '+' represent outlier data points that are beyond 1.5 * IQR.

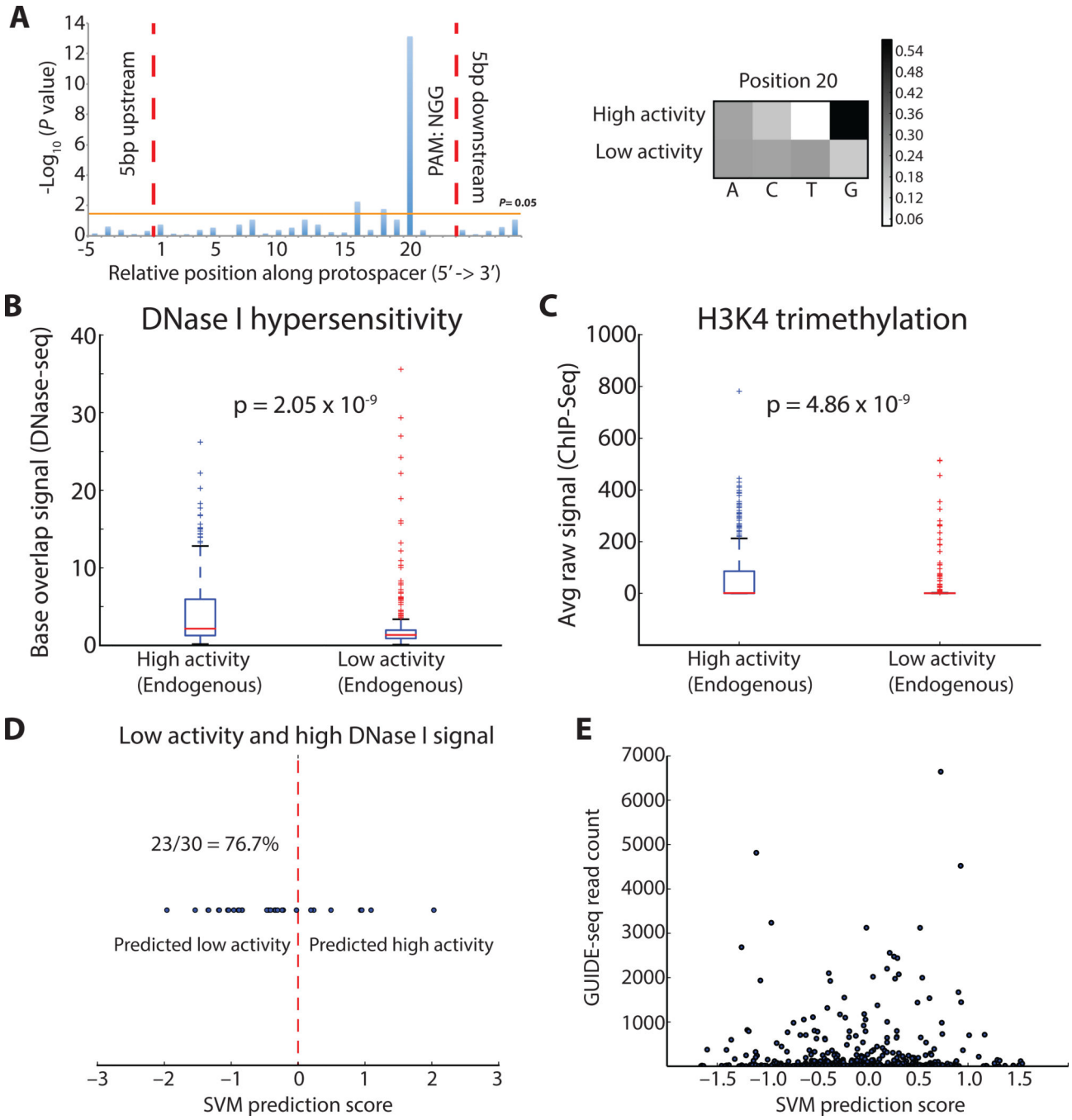


**Figure 3.**

**(A)** Position by position comparison of the base distributions between high and low activity sgRNA sequences. Position 20 exhibited the most striking difference. P values were calculated using a 2×4 Fisher's exact test. Comparison of **(B)** DNase I hypersensitivity and **(C)** H3K4-trimethylation between regions of high and low sgRNA activity. P-values were calculated by employing a t-test comparing the values from each group. **(D)** Using the top 30 sgRNA sequences corresponding to regions with low activity and high DNase I sensitivity, the sequences were scored by the SVM for predicted activity. Strikingly, the majority of the

sequences were predicted to have low activity. **(E)** Frequency of off-target activity and predicted activity. Using a recently published dataset, we scored all off-target sites with our SVM and compared those scores with the frequency of off-target activity observed. No discernable relationship was observed, suggesting these two aspects are independent.